Maximum Search Relevancy
Webmaster Best Practices

by Kevin Paddock, DTS Search Administrator

1. Use a Robots.txt file and the Robots meta-tag.

Search engines love to find a robots.txt file in the root of your web server. This file file contains a list of directories and/or files you want excluded from search engine spiders (crawlers). All reputable web crawlers honor your list -- as long as there are no syntax errors in it -- so exclude everything that does not contain information relevant to visitors, like cgi program directories; documents stored on web site folders but not ready for publication; new pages still being tested or awaiting approval; old, irrelevant documents; forms-based applications, etc. Crawler are very good at finding files that are not linked on any of your pages. Furthermore, for crawlers to grab unnecessary files is to waste your connection bandwidth which is better utilized for visitors navigating your site. Help search spiders get in and get out as fast as possible by using the robots.txt file. The structure of a robots.txt file looks like this:

a. Robots Tag: <META name="ROBOTS" content="NOFOLLOW,NOINDEX">
b. Robots.txt file format:

```
User-agent: *
Disallow: /calendar/
Disallow: /cgi/
```

Helpful robots.txt links can be found in the Appendix below.

2. Help content providers prepare documents for publication.

a. **Use a content management system.** The following data quality guidelines are more easily followed if you publish using a content management system (CMS). The importance of web content management cannot be underestimated or overemphasized.

b. **Put a unique Title on all your pages and documents.** "Welcome to *DepartmentName*" is a useless page title and "Microsoft Word" or "Untitled" are useless document titles too. Keep in mind, the title for your home page should indicate the purpose or identity of the entire site; the title for all other pages indicates the identity of the content on that page. Word and PDF documents contain a properties page that has a field named Title. This field is used by most search engines as the clickable title in search results. Content providers need to enter meaningful text in this field. Try this search on Google.com: allintitle:untitled document. You'll see anywhere from 60 to 90 million documents returned. You don't want search results like that from your site. Title length should be less than 80 characters. Avoid titles that consist of a single word.

c. **Populate the Description meta-tag** (html pages) and the Description property (Word and PDF documents). Relevancy is greatly enhanced if the Description tag has strong nouns, noun-phrases and verbs used to describe the content in greater detail. Use a maximum of 200 words. If your documents do not contain a Description, most search engines will provide one by extracting text from the content. Human-generated descriptions are almost always superior to search engine generated descriptions.

d. **Use well-structured, standard XHTML** in your page designs. Pages with substandard XHTML coding (missing tags, syntax errors, inaccurate titles, not to mention broken links) will score low on the relevancy scale. For example, use the structure tags (h1, h2, etc.) and the STRICT DTD to build your pages. A great way to attain this goal is to comply with the State of California Web Accessibility Guidelines based on the Federal Rehabilitation Act, Section 508, subpart 1194.22 and the W3C Web Content Accessibility Guidelines, (WCAG) level "double-A". You not only get 508 and WCAG complaint pages, you get search engine optimized pages!

e. **Validate all HTML content** Web browsers are fairly forgiving of mal-formed HTML, but not search engines. Use your browser developer tools or web-based validation to find code deficiencies. See Appendix for some good resources.

f. **Use the meta-Keywords tag**: It doesn't hurt to populate the content of this tag. Despite that fact that most Internet search engines ignore the keywords meta, the State of California search engine DOES look for it, so you will improve search relevancy if you use it. Provide your most important keywords and one or two synonyms; make sure the keywords you choose are indeed found in your documents. No more than 20 words; 10 words is ideal. DO NOT put the keywords in your templates! If every page has the same keywords, this meta tag is almost useless.

g. **Spell check EVERY document** before submitting it for publication. Many search engines do not use a dictionary to find misspelled words; they compare the terms entered with words (and non-words) in the collection indexes. The "Did you mean?" feature, then, depends on correct spellings in your documents. To see what I mean, search www.ca.gov for these non-words: "governer", "Californa", and "Californai". Rather embarrassing.

h. **Avoid Using Frames.** Frames do not fit the "conceptual model" of the web -- where every document corresponds to a single URL.

i. **Make web pages for users, not search engines.** The art of Search Engine Optimization (SOE) is of secondary importance to well-designed, information-rich, text-focused web sites. Be sure the ALT text for any non-text objects (Flash, images, etc.) is descriptive and any links to the content use strong descriptions in the anchor text.

3. Make your sites easy to navigate.

Establish a clear hierarchy of hypertext links. Every page should be reachable from at least one hypertext link. Offer a site map to all the important parts of the site. Google recommends using jump pages or basic ***html*** site maps. Jump pages are similar to site maps and offer a list of links that lead the user deeper within the site. The crawler will then be able to follow the links contained on these pages. Link to no more than 100 other pages from any page. Menus driven by Javascript should be avoided.

Maximum Search Relevancy
Webmaster Best Practices
by Kevin Paddock

4. Interlink (or cross-link) from other sites

Google's Page Rank algorithm ranks pages higher if they are linked to by other sites. This is often not an applicable technique for State Government web sites, but to the extent that you can cross-link your sites or other State sites, doing so ***will*** improve the quality of search results.

5. Validate all HTML content

Use Validation tools, web sites and a text browser to insure that your code is valid and accessible (section 508 and W3C compliant). See Appendix for some links.

6. Clean old files off your web servers.

If you aren't careful to remove old files with a robots.txt file (or the delete key), search crawlers will find them - even if they are not linked on any page! Some enterprises will NOT use the Google Search Appliance because it finds documents that may present a security compromise to the organization. Keep in mind, old data is usually irrelevant. Unless you have good reason to leave old data on your sites, remove the files or at the very least exclude them in the robots.txt file.

7. Follow File Size Limit guidelines.

No HTML pages over 2.5MB (that's huge!). No non-HTML (PDF, etc.) over 30MB. Files over these sizes will be indexed, but only up to these size limits. The rest of the file content will be ignored. Realistically, PDFs over 2MB are still a pain to download for most visitors; in March of 2006, 58% of home internet connections were still dial-up. Despite the visitor bandwidth, large PDFs (over 5MB) are considered bad form.

8. Get Listed.

To insure your sites are crawled for inclusion in state-wide searches, be sure they are listed in the State Agency Index. If they are not listed, we have no way of knowing your exact domain name to crawl. A link to the State of California Agency Index can be found on the California Homepage at www.ca.gov.

9. Register as a California Site.

Register a .ca.gov Domain Name. Fully-qualified State of CA domain names include the domain suffix ".ca.gov". If you have registered other top-level domains like ".org" or ".com", register a ".ca.gov" alias for you site as well. The only exception to this rule is educational institutions which are obviously registered under the top-level domain ".edu". By following this naming convention, you insure that the search spider identifies your web site as a genuine State of California site. Note: the domain suffix ".state.ca.us" is deprecated and should be avoided.

Maximum Search Relevancy
Webmaster Best Practices
by Kevin Paddock

10. Use a link checker to find broken links.

Crawlers waste lots of time retrying inaccessible pages from broken links. The more broken links the longer it takes to finish crawling your site. And for obvious other reasons, you don't want broken links, period.

11. Make Search-Engine Friendly URLs.

This best practice is very important and a bit technical. Background: database driven and dynamic sites present pages based on name value parameter pairs in the URL. You have such a site if your pages are generated by .asp, .jsp, .php, etc. scripts. Search engines -- including the Google Search Appliance -- have various degrees of difficulty with these pages. Many search engines skip the script files completely -- to avoid problems. If configured to index them, the GSA can get "stuck" in a veritable "black hole" of pages created with on-the-fly parameters. This tryst with your web site ends when the GSA hits some non-user configurable document threshold and moves on. Unfortunately, every document it finds -- most of these are login screen, data entry forms and empty faq pages -- counts against our paid-for document limit AND, search results for any actual words found on these items are mostly useless. Google for "search engine friendly URL" and see the Appendix below.

12. Remove Duplicates.

If you see duplicate entries in search results for your files, it may be because you actually have duplicates on your website. Remove one of the duplicates. Link checkers can help you find duplicates.

Appendix: Webmaster Resources

Helpful Resources for Webmasters:

Google Sitemap: https://www.google.com/webmasters/sitemaps/docs/en/about.html

Lynx Browser: http://csant.info/lynx.htm

Xenu's Link Sleuth: http://home.snafu.de/tilman/xenulink.html

HTTP Header Viewer: http://www.delorie.com/web/headers.html

Validation Tools: http://www.w3.org/QA/Tools/ or http://validator.w3.org/

Page Cleanup Tools: http://www.w3.org/People/Raggett/tidy/

Robots exclusion tutorials:
http://www.robotstxt.org/wc/exclusion.html and
http://www.outfront.net/tutorials_02/adv_tech/robots.htm
Robots syntax checker: http://tool.motoricerca.info/robots-checker.phtml
One little mistake and the entire list will not work. So syntax checking is very important.
Robots.txt file generator:
http://www.webtoolcentral.com/webmaster/tools/robots_txt_file_generator/

Search Engine Friendly URLs:
http://www.sitepoint.com/article/search-engine-friendly-urls
http://www.websitepublisher.net/article/search_engine_friendly_urls/

Comments: searchMaster@dts.ca.gov